3DV-TON: Textured 3D-Guided Consistent Video Try-on via Diffusion Models

Min Wei[1,2] Chaohui Yu[†1,2] Jingkai Zhou[1,2,3] Fan Wang[1]
[1]DAMO Academy, Alibaba Group [2]Hupan Lab [3]Zhejiang University,
{weimin.wei, huakun.ych}@alibaba-inc.com

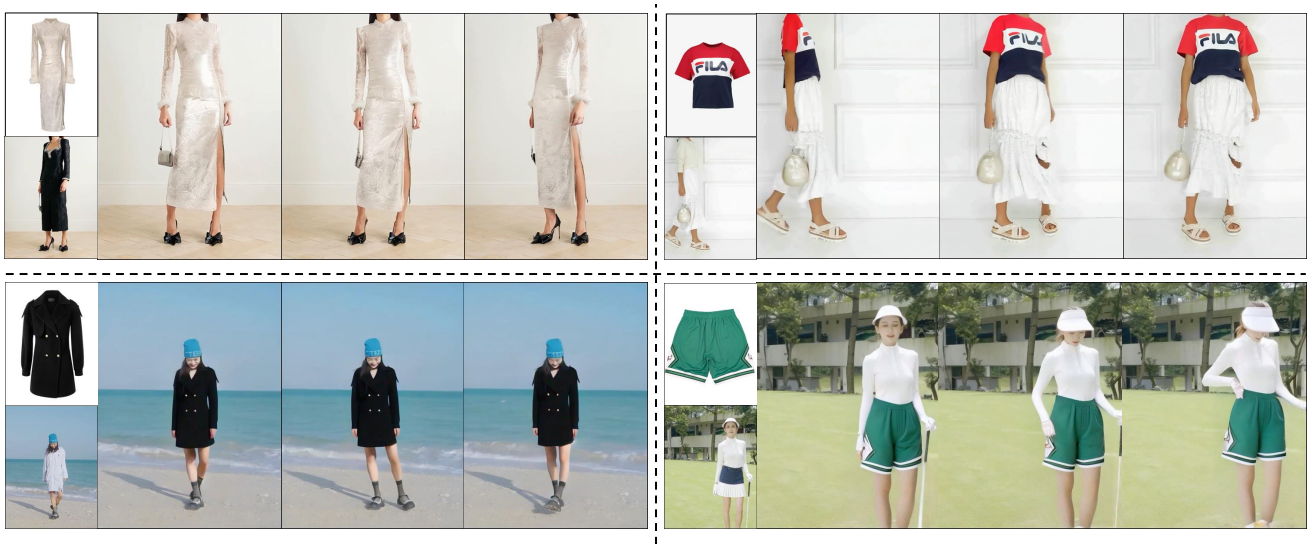

Figure 1. **Try-on videos generated by 3DV-TON.** Our method can handle various types of clothing and body poses, while accurately restoring clothing details and maintaining consistent texture motion.

Abstract

Video try-on replaces clothing in videos with target garments. Existing methods struggle to generate high-quality and temporally consistent results when handling complex clothing patterns and diverse body poses. We present 3DV-TON, a novel diffusion-based framework for generating high-fidelity and temporally consistent video try-on results. Our approach employs generated animatable textured 3D meshes as explicit frame-level guidance, alleviating the issue of models over-focusing on appearance fidelity at the expanse of motion coherence. This is achieved by enabling direct reference to consistent garment texture movements throughout video sequences. The proposed method features an adaptive pipeline for generating dynamic 3D guidance: (1) selecting a keyframe for initial 2D image try-on, followed by (2) reconstructing and animating a textured 3D mesh synchronized with original video poses. We further introduce a robust rectangular masking strategy that successfully mitigates artifact propagation caused by leaking clothing information during dynamic human and garment movements. To advance video try-on research, we introduce HR-VVT, a high-resolution benchmark dataset containing 130 videos with diverse clothing types and scenarios. Quantitative and qualitative results demonstrate our superior performance over existing methods. The project page is at this link https://2y7c3.github.io/3DV-TON/

1. Introduction

Video try-on aims to change the person's clothing in the given video to a target garment, enabling customers to visualize themselves wearing clothing items without physical trials through enhanced immersion and interactivity. The process must preserve intricate garment details while maintaining consistent texture representation throughout the video sequence.

Prior video try-on works [11, 30, 70] typically employ

† Corresponding author.

Figure 2. **Textured 3D guidance.** We construct the textured 3D guidance based on image try-on results, then animate the mesh after pasting the texture, providing a consistent texture motion reference on the appearance level.

flow-driven warping modules [5, 12, 16, 54, 58] for precise garment alignment on human figures, complemented by neural generators to synthesize the final appearance. However, these methods face inherent limitations from their reliance on warping operations: while effectively adapting garment geometry through shape deformation to match pose variations, they inherently compromise temporal coherence in generated sequences. This fundamental constraint hinders handling of substantial clothing deformations and complex occlusions, limiting practical application to simplified scenarios.

Recent advancements [13, 62] harness pre-trained diffusion models [27, 49] to address limitations of conventional warping modules. These works [13, 62] implement a dual-UNet architecture: a primary denoising UNet [50] alongside a parallel reference UNet that directly extracts garment features, eliminating explicit warping. Hierarchical temporal attention layers [20] are integrated within the denoising net to model motion dynamics and mitigate inter-frame inconsistencies. Concurrently, Diffusion Transformer (DiT)-based frameworks [46] demonstrate enhanced performance in video try-on through superior generative scalability, as evidenced by works like [8, 69]. Nevertheless, empirical analysis in [4] reveals that pixel-reconstruction objectives in video diffusion models remain constrained in achieving robust temporal coherence.

In this paper, we present 3DV-TON, a diffusion-based framework for generating high-fidelity temporally-consistent video try-ons. To tackle the limitation, models prioritize appearance fidelity over motion coherence, in prior literature where pixel-based reconstruction objectives inherently, we introduce explicit frame-level textured 3D guidance. Our method directly models 3D human meshes wearing target garments, ensuring spatiotemporal consistency across diverse poses and viewpoints through motion-aligned mesh propagation, which providing a consistent motion reference on the appearance level. While existing methods [71, 72] employ 3D human priors, they exclusively utilize geometric structural cues without textured guidance. Our experiments demonstrate that geometric-only guidance (e.g., SMPL [41, 45]) often fails to sufficiently constrain models, resulting in appearance-biased optimization and motion artifacts. Crucially, our textured 3D guidance uniquely preserves garment identity throughout video sequences, addressing a critical oversight in current video try-on works.

As illustrated in Figure 2, our pipeline begins with selecting a frame through pose estimation, processed using advanced diffusion-based image try-on methods [6, 7, 61]. This initial frame undergoes animatable textured 3D mesh reconstruction aligned with the source video motion to generate temporally consistent reference sequences. Unlike the previous warp module, our framework leverages single-image 3D reconstruction [48, 59, 60, 67] to inherently establish spatiotemporal consistency, delivering robust appearance priors for the denoising UNet while reducing temporal attention dependencies. This strategy effectively bypasses complex warping operations through mesh animation, while benefiting from mature single-image reconstruction methods without task-specific retraining.

We further propose a dynamic rectangular masking strategy to prevent garment information leakage during human motion, which is a primary failure source in video try-on. To counter excessive masking, we implements both clothing images and try-on images as references to provide garments and environment context, and design an effective guidance feature extraction and fusion diffusion-based architecture. Comprehensive experiments demonstrate that our 3D-aware framework achieves superior visual quality and consistency in complex dynamic scenarios compared to existing approaches.

In summary, our main contributions are as follows:

- We propose 3DV-TON, a novel diffusion-based video try-on method that employs textured 3D guidance to alleviate motion incoherence stemming from appearance bias. Our method effectively generates try-on videos maintaining consistent texture motion across varying body poses and camera viewpoints.
- We introduce a 3D guidance pipeline capable of adaptively generating animatable textured 3D meshes, ensuring consistent texture guidance across both spatial and temporal domains. The framework seamlessly integrates with existing methodologies without necessitating additional training.
- We establish a high-resolution video try-on benchmark enabling better evaluation of recent works, and demonstrate that our 3DV-TON outperforms existing video try-on methods in both quantitative and qualitative experiments.

2. Related Works

Image Virtual Try-on. Image virtual try-on aims to generate images of a target person wearing a given clothing. Many GAN-based methods [5, 12, 16, 24, 29, 54, 58, 65] typically first warp the clothing image onto the target person's body. Then, a generator is used to blend the warped clothing with the human body to produce realistic results. These methods rely on the accuracy of the warping module. Due to undesired distortions and artifacts caused by the TPS [1]-based methods [21, 43], many subsequent methods [16, 22, 24, 29, 58] have focused on predicting dense flow to achieve better warping of clothing, and have made significant progress. However, explicit warping techniques still struggle with complex poses and occlusions.

Recently, several works [6, 7, 18, 33, 44, 61] tend to employ powerful pre-trained diffusion models [27, 49] as an alternative to GANs [17] to generate more realistic try-on results. OOTDiffusion and IDM-VTON [6, 61] utilized a dual-UNet [50] structure and integrated clothing features and person features through self-attention. CatVTON [7] proposed to merge dual-UNet architectures, simplifying the training parameters and the inference process. However, applying image-based try-on techniques frame by frame to videos can lead to temporal inconsistent results.

Video Virtual Try-on. Compared to image try-on, video try-on needs to maintain temporal consistency between frames to generate realistic, high-quality results, which adds more challenges to the task. Previous works [11, 30, 70] typically employs a flow-based warping module [5, 12, 16, 54, 58] for precise garment alignment on human bodys, and combine the warped clothing with the person in the video. In video try-on, warp-based methods also face challenges in handling complex textures and motion.

Recent diffusion-based works [13, 25, 55, 62], build on the dual-UNet architecture, a primary denoising UNet [50] alongside a parallel reference UNet that directly extracts garment features to preserve the visual quality, and insert hierarchical temporal modules [20] to ensure temporal smoothness. ViViD [13] releaseed a new dataset and improves the generation resolution from 256 to 512. Some works [55, 62] utilized private datasets with a resolution of 512 and introduced techniques to emphasize the clothing. More recently, some works [8, 69] utilized the powerful diffusion transformer (DiT) framework and have made significant progress in the video try-on task. However, these methods struggle to maintain consistent temporal coherence between frames, tend to generate over-smoothed deformed clothing textures.

Clothed 3D Human Reconstrction. Previous works [10, 15, 39, 68] typically requires modeling a 3D human body model and a clothing model, and then fit the clothing onto the human body model. Additionally, when animate the model, physical simulation is introduced to generate natural clothing movement. One line of research, *e.g.*, DiffAvatar [39], proposed a methods for body shape and garment assets recovery from 3D scan of a clothed person, and utilized differentiable simulation for co-optimizing garment and human body. Such methods have very high requirements for the input data. On the other hand, several methods reconstructs a clothed human from a single image and simultaneously models the clothing along with the person for animation. ICON [59] used body-based normal estimation for implicit 3D reconstruction. ECON [60] significantly improved reconstruction robustness by integrating explicit shape-based approaches with normal priors. SIFU [67] proposed a side-view conditioned implicit function to achieve more accurate reconstruction results. More recently, some works [48, 57] introduced large reconstruction models (LRMs) to enable feed-forward clothed human reconstructions. These works are capable of generating photorealistic, animatable human avatars in seconds, but they struggle to produce flexible clothing motions.

3. Method

The overview pipeline of our 3DV-TON is illustrated in Figure 3. We first introduce the textured 3D guidance generation pipeline in Section 3.1. Then, the model architecture and training strategy are illustrated in Section 3.2.

3.1. Animatable Textured 3D Guidance

SMPL&SMPLX. The Skinned Multi-Person Linear (SMPL) model [41] is a 3D parametric human model that defines the shape topology of body. It uses shape parameters $\beta \in \mathbb{R}^{10}$ and pose parameters $\theta \in \mathbb{R}^{24\times 3}$ to represent the 3D human body mesh $M(\beta, \theta)$ as:

$$\begin{aligned} T_p(\beta,\theta) &= \overline{T} + B_s(\beta) + B_p(\theta), \\ M(\beta,\theta) &= W(T_p(\beta,\theta), J(\beta), \theta, \mathcal{W}), \end{aligned} \tag{1}$$

where $\overline{T}$ is the mean template shape, $B_s(\beta), B_p(\theta)$ are vectors of vertices representing offsets from the template. $T_p(\beta,\theta)$ is the non-rigid deformation from $\overline{T}$. $W(\cdot)$ is the linear blend skinning (LBS) [53] function applied to rotate the vertices around the joint center $J(\beta)$ with the smoothing defined by the blend weights $\mathcal{W}$.

The SMPL-X model [45] builds upon SMPL, adding features for hands and face, enhanceing facial expressions, finger movements, and detailed body poses.

Clothed Human Reconstruction&Animation. Our human reconstruction method is based on ECON [60]. Given a video, we choose a frame I according to estimated body pose adaptively, which performing image try-on [6, 7, 61] during inference, as the input of normal estimation network [32, 59, 60]. To guide the normal map prediction for clothed normal map (denoted as $\hat{\mathcal{N}}^c_{\{F,B\}}$ where F, B denote front/back view), and ensure robustness across poses,

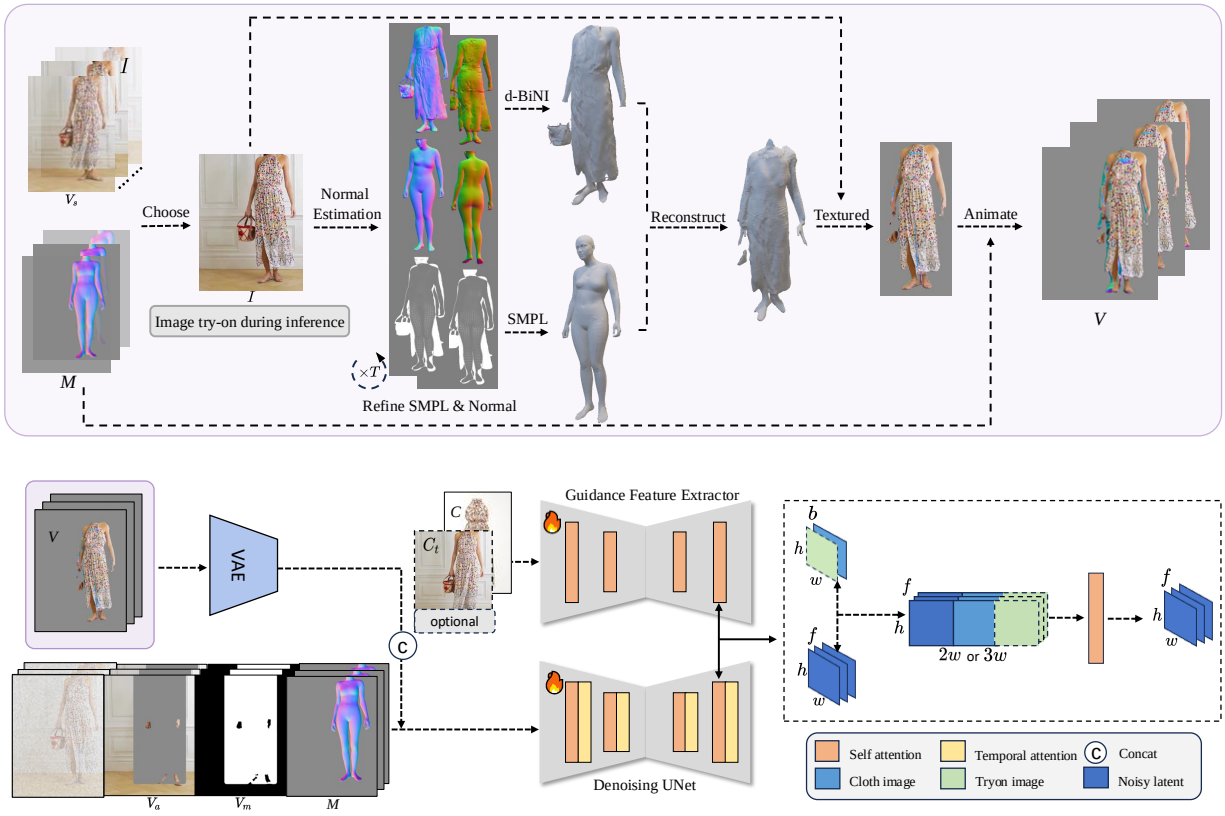

Figure 3. **The overview of 3DV-TON.** Given a video, we first use our 3D guidance pipeline to select a frame I adaptively, then reconstruct a textured 3D guidance and animate it align with the original video, *i.e.* V. We employ a guidance feature extractor for the clothing image C and the try-on images C_t, and perform feature fusion using the self-attentions in the denoising UNet.

we use body normal maps $\mathcal{N}^b_{\{F,B\}}$ rendered from the estimated SMPL-X $M^b(\beta, \theta)$ as reconstruction conditions. Accurate alignment between body estimation and clothing silhouettes proves crucial for this process. However, existing human pose and shape (HPS) regressors [36, 37, 51, 52] fails to provide pixel-aligned SMPL-X fits. Unlike previous works [59, 60] that require precise body pose optimization, our method prioritizes clothing reconstruction accuracy over anatomical details. By eliminating SMPL-X pose θ optimization during parameter refinement, we reduce optimization steps and reconstruction time to $\sim$30s while maintaining performance. We additionally optimize camera scale s to address systematic camera estimation errors in HPS methods. Our optimization process initializes with estimated SMPL-X's shape β, translation t parameters and camera scale s, focusing on minimizing silhouette and normal loss:

$$\mathcal{L}_{\text{SMPL-X}} = \mathcal{L}_{N_d} + \mathcal{L}_{S_d} + \lambda \cdot \min(\mathbf{d} - s, 0), \tag{2}$$
$$\mathcal{L}_{N_d} = |\hat{\mathcal{N}}^c - \mathcal{N}^b(\beta, t, s)|, \quad \mathcal{L}_{S_d} = |\hat{\mathcal{S}}^c - \mathcal{S}^b(\beta, t, s)|,$$

where $\mathcal{L}_{N_d}$ is a normal map L1 loss, $\mathcal{L}_{S_d}$ is a L1 loss between the silhouettes of the SMPL-X $\mathcal{S}^b$ and the clothed human mask $\hat{\mathcal{S}}^c$ segmented from image I. We additionally introduce a unidirectional regularization penalty to address frequent partial body observations in training data (see *Supplementary Materials*.), activated during loss computation when camera scale falls below the dataset-defined threshold $\mathbf{d}$. Following SMPL-X refinement, we iteratively updating the normal map and SMPL-X parameters through T refinement cycles.

We reconstruct the front and back surface using depth-aware silhouette-consistent bilateral normal integration (d-BiNI) method introduced by [2, 60]. However, poses often result in self-occlusions, which cause large portions of the surfaces to be missing. In such cases, we use a simple way to infill the missing surface using the estimated SMPL-X body that invisible to front or back cameras, and union the parts of surface by surface reconstruction methods [28, 31]. Since the reconstructed mesh is aligned with the image pixels, we can simply use interpolated pixel values as the mesh texture after calculating visibility. For the invisible body areas, we use normals as the texture.

The reconstructed clothed human inherit the hierarchical skeleton and skinning weights from the underling SMPL-

X body model, allowing to animate it using the estimated SMPL poses [51] from the original video. Specifically, for each vertices j of the clothed human mesh, we use k-nearest neighbor (KNN) search to obtain a set $\mathcal{K}_j$ composed of K neighboring control points denoted as $\{p_k | k \in \mathcal{K}_j\}$ in canonical SMPL-X model. Then, the interpolation weights for control points p_k can be computed as:

$$w_{jk} = \frac{\hat{w}_{jk}}{\sum_{k \in \mathcal{K}_j} \hat{w}_{jk}}, \quad \hat{w}_{jk} = \exp(-d_{jk}^2), \tag{3}$$

where d_{jk} is the distance between vertices j and the neighboring vertices p_k in SMPL-X. The overall 3D guidance generation pipeline is depicted in the upper part of Figure 3.

3.2. Network Architecture

Controlled Diffusion Model. Stable Diffusion [27, 49] is the basis for our network that consists of a varialtional autoencoder (VAE) [34] and a denoising UNet [50]. Given an image x_0 and a control condition $\mathbf{c}$, the VAE first encodes the image x_0 into latent space: $z_0 = \mathcal{E}(x_0)$. The UNet learns to predict a noise ϵ_θ or velocity v_θ based on the control condition $\mathbf{c}$ and the noisy latent z_t : $z_t = \alpha_t z_0 + \sigma_t \epsilon$. The training loss of the UNet can be formulated as:

$$\mathcal{L}_{LDM} = \mathbb{E}_{z,c,\epsilon,t}[\|v_t - v_\theta(z_t, t, \mathbf{c})\|_2^2], \tag{4}$$

where t represent the diffusion timestep, $\epsilon \sim \mathcal{U}(0, I)$, $v_t = \alpha_t \epsilon - \sigma_t z_0$ [40]. In inference, data samples can be generated from Gaussian noise $z_T \sim \mathcal{N}(0, I)$ by the denoising process.

Guidance Feature Extractor. Our method employs two reference conditions: clothing images $C \in \mathbb{R}^{b \times 3 \times H \times W}$ and try-on images $C_t \in \mathbb{R}^{b \times 3 \times H \times W}$ encoded into latent space through VAE encoder $\mathcal{E}$ as $\mathbf{C} = \mathcal{E}(C)$ and $\mathbf{C_t} = \mathcal{E}(C_t)$. These latent representations ($\mathbf{C}, \mathbf{C_t} \in \mathbb{R}^{b \times 4 \times h \times w}$ are concatenated along the batch dimension to form composite reference features $\mathbf{F} \in \mathbb{R}^{2b \times 4 \times h \times w}$. We duplicate the denoising UNet as the Guidance Feature Extractor that capture the visual features of the clothing images and try-on images. Note that we remove text encoders and all cross attention layers cause our textured 3D guidance provided sufficiently explicit visual reference.

Denoising Network. We employ a UNet architecture from Stable Diffusion [49] without cross-attention layers, extended into a pseudo-3D structure through temporal module [20] integration to enable realistic motion generation, serving as our base denoising network. Given a batch of source videos $V_s \in \mathbb{R}^{b \times 3 \times f \times H \times W}$, with corresponding clothing-agnostic videos $V_a \in \mathbb{R}^{b \times 3 \times f \times H \times W}$ and mask videos $V_m \in \mathbb{R}^{b \times 1 \times f \times H \times W}$. We extimate the SMPL sequences M using HPS methods [51, 52]. Our adaptive 3D guidance pipeline then generates textured 3D guidance V. The denoising input comprises concatenated features along the channel dimension: the noisy latent video z_t, the latent clothing-agnostic video $\mathcal{E}(V_a)$, the resized mask video V_m', the SMPL geometic guidance $\mathcal{E}(M)$ and the textured 3D guidance $\mathcal{E}(v)$. To accommodate this 17-channel input, we expand the UNet's initial convolutional layer with zero-initialized weights.

Our guidance feature extractor avoids feature fusion between clothing and try-on images. Instead, we implement texture-aware fusion through spatial attention mechanisms (Figure 3). For each latent $x_s^i \in \mathbb{R}^{b \times c \times f \times h \times w}$ entering the i-th self-attention layer, we retrieve corresponding reference features corresponding reference feature $x_f^i \in \mathbb{R}^{2b \times c \times h \times w}$ from the extractor. These features split into clothing feature x_c^i and the try-on feature $x_{c_t}^i$, which we temporally align by replicating along the frame dimension to obtain $\hat{x}_c^i, \hat{x}_{c_t}^i \in \mathbb{R}^{b \times c \times f \times h \times w}$. As shown in Figure 3, the three types of features are concatenated along the spatial dimensions, denoted by $\hat{x}_s^i \in \mathbb{R}^{b \times c \times f \times h \times 3w}$, . Then feature fusion is performed by the attention layer of the denoising network to obtain the latent x_s^{i+1}, which incorporates both the fine clothing textures and the frame-consistent 3D features.

Training Strategy. Inspired by [13, 47, 63], our model is trained on both image and video datasets by treating images as single-frame videos. During training, we randomly select a type of dataset via a random number $r \sim \mathcal{U}(0, 1)$, where $\mathcal{U}(\cdot, \cdot)$ is the uniform distribution. If $r < \tau$, we use the sampled images for training, and set the gradients of the temporal attention as zero to freeze the temporal module. Otherwise, we sample data from the video dataset, and make the temporal attention trainable. Hence, our training objective can be formulated as:

$$\mathcal{L} = \mathbb{E}_{z,\epsilon,t}[\|v_t - v_\theta(z_t, t, \mathbf{C}, \mathbf{C_t}, \mathbf{V})\|], \; z \sim \{z_{img}, z_{vid}^{1:f}\}. \tag{5}$$

We incorporate control conditions through Classifier-Free Guidance (CFG) [26]. Specifically, we randomly omit the clothing image $\mathbf{C}$ with a probability of p_1, the try-on image $\mathbf{C_t}$ with a probability of p_2, and the textured 3D guidance $\mathbf{V}$ with a probability of p_3.

Masking Strategy. Our pipeline begins with garment segmentation using either human parsing [38] or segmentation model [35] to generate clothing masks. We compute bounding boxes from these masks and employ a human estimation model [19, 36, 37]to selectively critical anatomical regions (*e.g.* face and hands) while preserving body detail. This streamlined approach effectively prevents garment transfer failures caused by leaking clothing. Please refer to Section 4.5 for illustration.

Figure 4. **Qualitative comparison for dress try-on on the ViViD dataset.**

Method	Paired				Unpaired	
	SSIM↑	LPIPS↓	$VFID_{I3D}$↓	$VFID_{ResNeXt}$↓	$VFID_{I3D}$↓	$VFID_{ResNeXt}$↓
StableVITON [33]	0.8019	0.1338	34.2446	0.7735	36.8985	0.9064
OOTDiffusion [61]	0.8087	0.1232	29.5253	3.9372	35.3170	5.7078
IDM-VTON [6]	0.8227	0.1163	20.0812	0.3674	25.4972	0.7167
StableVITON+AM [63]	0.8207	0.1291	19.9239	0.7586	22.0262	0.8283
OOTDiffusion+AM [63]	0.8154	0.1244	19.3173	0.9382	23.3938	1.1485
IDM-VTON+AM [63]	0.8252	0.1212	18.2048	0.4481	22.5881	0.5397
ViViD [13]	0.8029	0.1221	17.2924	0.6209	21.8032	0.8212
CatV²TON [8]	0.8727	0.0639	13.5962	0.2963	19.5131	0.5283
3DV-TON (Ours)	0.8681	0.0707	13.4062	0.2741	19.4714	0.3664
*3DV-TON** (Ours)	**0.8992**	**0.0521**	**10.9680**	**0.2033**	**18.1151**	**0.3149**

Table 1. **Quantitative comparison on the ViViD dataset.** * indicates our method using the same mask with ViViD [13].

4. Experiments

4.1. Datasets

Training datasets. We use two image datasets, VITON-HD[5] and DressCode [9], along with one video dataset, ViViD [13], to train our diffusion model. Due to the low resolution of the VVT [11] dataset, we opt not to use it for training. Specifically, the VITON-HD dataset contains 13,678 images of upper-body clothing with corresponding model images. The DressCode dataset includes 15,363 images of upper-body clothing, 8,951 images of lower-body clothing, and 2,947 images of dresses, along with images of models wearing these garments. The ViViD dataset consists of 9,700 videos featuring models along with corresponding clothing images. This dataset contains 4,823 video-image pairs for upper-body clothing, 2,133 for lower-body clothing, and 2,744 for dresses, with a total of 1,213,694 frames. All image and videos are resized to 768×576 for training. For video data, we randomly select a frame to construct the try-on image condition using the image try-on method [6, 7, 61]. For image data, we set all try-on conditions to be empty.

HR-VVT benchmark. Owing to the limitations of the ViViD dataset, which contains limited scenarios, and VVT dataset, which only includes upper-body clothing and exhibits relatively uniform body poses, coupled with a low resolution of only 256×192, it is challenging to accurately assess video try-on methods. Therefore, we have constructed a high-resolution ($\sim$720p) video try-on benchmark called HR-VVT that includes 130 videos with 50 upper-body clothing, 40 lower-body clothing, and 40 dresses, with a variety of garments and motions in complex scenarios. Please refer to our *Supplementary Materials* for more dataset details.

4.2. Implementation Details

Textured 3D guidance. During the 3D human reconstruction process, to achieve better body estimation, we use a image-based HPS regressor [14, 37, 64] with SMPL-X and iteratively refined the SMPL-X parameters $T = 10$ times for clothed human reconstruction based on ECON [60]. To ensure smooth animation of the 3D guidance, we employ a video-based SMPL estimation approach [51, 52]. Due to the differences between image-based and video-based estimation methods, we use the body from the video-based estimation as the binding template to avoid texture distortion and animate it to render a guidance video, which is aligned with the source video. Please refer to *Supplementary Materials* for more details.

Training. We initialize our guidance feature extractor and denoising network using the weights from SD1.5 [49], and employ Animatediff [20] to initialize the temporal attention. Our model is trained in a single-stage manner using 768×576 resolution, 32 frames (2 strides for videos) data. The model was trained using A800 GPUs for 40000 steps with a learning rate of 1e-5.

4.3. Qualitative Results

We conduct qualitative comparisons with the currently available method for video try-on that released the inference code, ViViD [13] and CatV^{2}TON[8]. The clothing images and the person videos are from ViViD-S [8, 13] test set and our HR-VVT set. None of the images or videos have appeared in the training data.

Comparisons on ViViD dataset. As shown in Figure 4 and Figure 5, other methods suffer from artifacts and generate garments limited by the patterns of the original clothing, while our method generates accurate garment shapes, offers better visual quality, and produces realistic clothing motion that adapts to the person's motions. Figure 6 shows that

Figure 5. **Qualitative comparison for upper garment try-on on the ViViD dataset.**

Figure 6. **Comparison for lower garment try-on on ViViD.**

ViViD [13] fails at this case, while CatV^{2}TON generates inconrrect gaments along with blurriness and artifacts. In contrast, Our 3DV-TON generates accurate clothing with good temporal consistency.

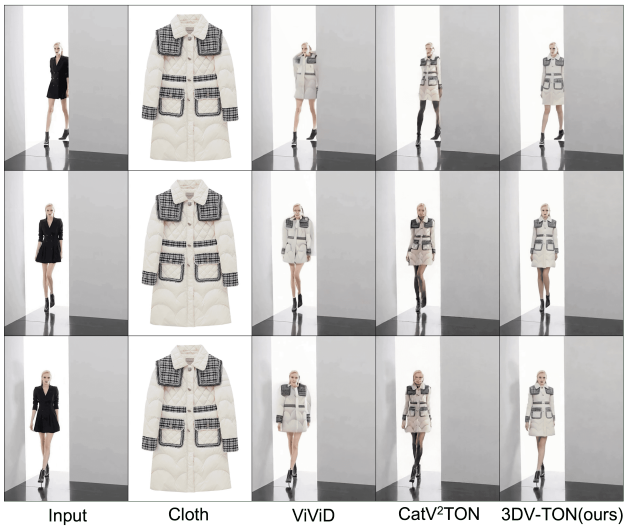

Figure 7. **Qualitative comparison for dress try-on on HR-VVT.**

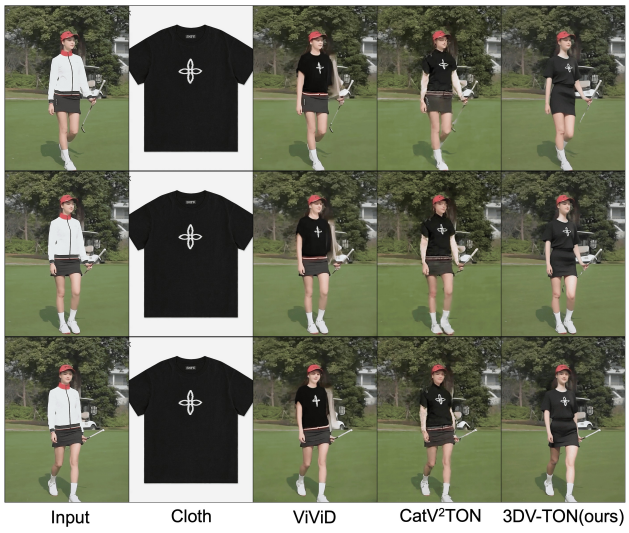

Figure 8. **Comparison on upper garment try-on on HR-VVT.**

Comparisons on HR-VVT benchmark. Our HR-VVT benchmark includes a more diverse set of environments, clothing. As shown in Figure 7, ViViD [13] struggles with perspective changes during subject movement, and CatV^{2}TON fails to preserve garment consistency. In contrast, our method leverages explicit textured 3D guidance to maintain visual coherence across viewpoints and motion sequences. Figure 8 shows that our approach's superiority in outdoor scenarios, where competing methods exhibit artifacts and unrealistic texturing. Figure 9 demonstrates how our robust 3D guidance pipeline ensures reliable per-

Figure 9. **Comparison on lower garment try-on on HR-VVT.**

Method	SSIM↑	LPIPS↓	Paired $VFID_{I3D}$↓	Paired $VFID_{ResNeXt}$↓	Unpaired $VFID_{I3D}$↓	Unpaired $VFID_{ResNeXt}$↓
ViViD [13]	**0.8889**	0.0876	**10.2367**	0.1785	16.4684	0.6807
CatV²TON [8]	0.8670	0.1144	12.1280	0.1798	16.8880	**0.3454**
3DV-TON (Ours)	0.8801	**0.0857**	10.7682	**0.1420**	**14.5499**	0.4217

Table 2. **Quantitative comparison on HR-VVT benchmark.** Best results are highlighted in bold, the second are underlined.

Datasets	Method	Fidelity (%)	Consistency (%)	Overall Quality (%)
ViViD	ViViD [13]	24.55	20.25	20.39
	CatV²TON [8]	12.09	10.92	10.53
	3DV-TON (Ours)	**63.36**	**68.83**	**69.08**
HR-VVT	ViViD [13]	14.02	11.82	11.77
	CatV²TON [8]	5.97	3.36	2.70
	3DV-TON (Ours)	**80.01**	**84.82**	**85.53**

Table 3. **User preference rate on the HR-VVT benchmark and ViViD dataset.**

formance even with partial character visibility. Please refer to our *Project Page* for more qualitative comparisons and video results.

4.4. Quantitative Results

Comparisons on ViViD dataset. We report quantitative results with SSIM [56], LPIPS [66] to evaluate the image visual quality in the paired setting, and use Video Frechet Inception Distance(VFID) [11, 42] to measure the generation quality and temporal consistency in the both paired and unpaired setting, following [8, 11, 13, 30]. VFID extracts features of video clips for computation using pretrained video backbone I3D [3] and 3D-ResNeXt101 [23]. Our method employs a rectangular mask strategy that enlarges the area to be generated, which creates an unfair comparison. Nonetheless, as reported in Table 1, our method still achieves comparable results in SSIM and LPIPS metrics, while surpassing existing methods in the VFID metric.

Figure 10. **Ablations for the mask strategy.**

Figure 11. **Ablations for the SMPL guidance.**

When we use the mask from ViViD [13], our method delivers better results across all metrics.

Comparisons on HR-VVT benchmark. We compare the current state-of-the-art and code released video try-on method, ViViD [13] and CatV^{2}TON on our benchmark. As shown in Table 2, although we use the larger mask, our method outperforms other works. This improvement can be attributed to the consistent texture features brought by our textured 3D guidance. We also demonstrated advantages in LPIPS, which proves that our method is capable of generating try-on results with better visual quality.

User Study. Considering that the current quantitative metrics are difficult to accurately evaluate the quality of the model in terms of human preference in the unpaired setting without ground truth. We conduct a user study that includes 130 video results and involved 20 annotators to provide a comprehensive comparison in terms of visual quality and motion consistency. Table 3 shows that our 3DV-TON achieves better motion coherence and effectively restores clothing details (*i.e.* "Fidelity"), resulting in superior visual quality.

4.5. Ablation Study

We conducted ablation studies to verify the effectiveness of our textured 3D guidance.

Speed Analysis. After optimizing the SMPL fitting process, our method is capable of completing the reconstruction in ∼30s, and generating a 32-frame video with dif-

Figure 12. **Ablations for the textured 3D guidance.** Textured 3D guidance helps to improve the motion conherence.

Figure 13. **Ablations for our textured 3D guidance.** Textured 3D guidance helps to improve the clothing consistency.

SMPL	Tex. 3D	SSIM↑	LPIPS↓	$VFID_{I3D}$↓	$VFID_{RexNeXt}$↓
		0.858	0.078	5.236	1.0257
✓		0.880	0.059	4.087	0.5854
✓	✓	**0.909**	**0.048**	**2.381**	**0.3011**

Table 4. **Quantitative ablations for the 3D guidance.**

fusion after removing cross attention takes ∼35s under 768×576 resolution. Due to our use of single-image reconstruction, the diffusion model accounts for the majority of the inference time for longer videos.

Mask Strategy. Our robust rectangular mask strategy can effectively addresses the issue of try-on failures caused by the leakage of original clothing information in videos. Figure 10 demonstrates that our method can generate try-on results that align more closely with the target garment patterns.

SMPL Guidance. As shown in Figure 11, the introduction of SMPL guidance helps in generating more accurate human bodies and properly fit clothing on the body. The person's arms and shoulders are accurately generated after using SMPL.

Textured 3D Guidance. Recent studies [4] demonstrate that conventional pixel reconstruction objective biases diffusion models toward appearance fidelity while compromising geometric accuracy, leading to motion artifacts. As demonstrated in Figure 12, while SMPL-based geometric guidance improves body structure estimation in masked regions, it exhibits persistent limb ambiguity during leg-crossing scenarios. Our textured 3D guidance resolves this limitation by supplementing explicit appearance constraints, effectively balancing visual quality and motion coherence. Our texture 3D guidance ensures accurate clothing texture preservation across arbitrary poses and viewpoints. As shown in Figure 13 (a), our method faithfully reconstructs the "boss" logo while maintaining anatomically consistent body proportions during lateral rotation. Figure 13 (b) demonstrates viewpoint-consistent rendering of the "lee" text across dynamic poses. In Table 4, we present quantitative ablation experiments, where geometric features and textured 3D guidance significantly improved the SSIM, LPIPS, and VFID metrics.

5. Conclusion

In this paper, we propose 3DV-TON, a novel diffusion-based framework guided by geometric and textured 3D guidance. By leveraging SMPL as parametric body geometry and employing single-image reconstructed 3D humans as animatable textured 3D guidance to provide frame-specific appearance conditions, 3DV-TON alleviates the critical limitation of inconsistent results caused by existing methods' over-focus on appearance fidelity. The framework learns a geometrically plausible human body across diverse poses and viewpoints, while maintaining temporally consistent motion of clothing textures. Quantitative and qualitative evaluations on existing datasets and our newly introduced HR-VVT demonstrate state-of-the-art performance in the video try-on task.

References

[1] Fred L Bookstein and WDK Green. A thin-plate spline and the decomposition of deformations. *Mathematical Methods in Medical Imaging*, 2(14-28):3, 1993.

[2] Xu Cao, Hiroaki Santo, Boxin Shi, Fumio Okura, and Yasuyuki Matsushita. Bilateral normal integration. In *European Conference on Computer Vision*, pages 552–567. Springer, 2022.

[3] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[4] Hila Chefer, Uriel Singer, Amit Zohar, Yuval Kirstain, Adam Polyak, Yaniv Taigman, Lior Wolf, and Shelly Sheynin.

Videojam: Joint appearance-motion representations for enhanced motion generation in video models. *arXiv preprint arXiv:2502.02492*, 2025.

[5] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14131–14140, 2021.

[6] Yisol Choi, Sangkyung Kwak, Kyungmin Lee, Hyungwon Choi, and Jinwoo Shin. Improving diffusion models for virtual try-on. *arXiv preprint arXiv:2403.05139*, 2024.

[7] Zheng Chong, Xiao Dong, Haoxiang Li, Shiyue Zhang, Wenqing Zhang, Xujie Zhang, Hanqing Zhao, and Xiaodan Liang. Catvton: Concatenation is all you need for virtual try-on with diffusion models. *arXiv preprint arXiv:2407.15886*, 2024.

[8] Zheng Chong, Wenqing Zhang, Shiyue Zhang, Jun Zheng, Xiao Dong, Haoxiang Li, Yiling Wu, Dongmei Jiang, and Xiaodan Liang. Catv2ton: Taming diffusion transformers for vision-based virtual try-on with temporal concatenation. *arXiv preprint arXiv:2501.11325*, 2025.

[9] Morelli Davide, Fincato Matteo, Cornia Marcella, Landi Federico, Cesari Fabio, and Cucchiara Rita. Dress code: High-resolution multi-category virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[10] Luca De Luigi, Ren Li, Benoit Guillard, Mathieu Salzmann, and Pascal Fua. Drapenet: Garment generation and self-supervised draping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1451–1460, 2023.

[11] Haoye Dong, Xiaodan Liang, Xiaohui Shen, Bowen Wu, Bing-Cheng Chen, and Jian Yin. Fw-gan: Flow-navigated warping gan for video virtual try-on. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1161–1170, 2019.

[12] Haoye Dong, Xiaodan Liang, Yixuan Zhang, Xujie Zhang, Xiaohui Shen, Zhenyu Xie, Bowen Wu, and Jian Yin. Fashion editing with adversarial parsing learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8120–8128, 2020.

[13] Zixun Fang, Wei Zhai, Aimin Su, Hongliang Song, Kai Zhu, Mao Wang, Yu Chen, Zhiheng Liu, Yang Cao, and Zheng-Jun Zha. Vivid: Video virtual try-on using diffusion models. *arXiv preprint arXiv:2405.11794*, 2024.

[14] Yao Feng, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, and Michael J Black. Collaborative regression of expressive bodies using moderation. In *2021 International Conference on 3D Vision (3DV)*, pages 792–804. IEEE, 2021.

[15] Daiheng Gao, Xu Chen, Xindi Zhang, Qi Wang, Ke Sun, Bang Zhang, Liefeng Bo, and Qixing Huang. Cloth2tex: A customized cloth texture generation pipeline for 3d virtual try-on. *arXiv preprint arXiv:2308.04288*, 2023.

[16] Yuying Ge, Yibing Song, Ruimao Zhang, Chongjian Ge, Wei Liu, and Ping Luo. Parser-free virtual try-on via distilling appearance flows. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8485–8493, 2021.

[17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[18] Junhong Gou, Siyu Sun, Jianfu Zhang, Jianlou Si, Chen Qian, and Liqing Zhang. Taming the power of diffusion models for high-quality virtual try-on with appearance flow. In *Proceedings of the 31st ACM International Conference on Multimedia*, 2023.

[19] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7297–7306, 2018.

[20] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *International Conference on Learning Representations*, 2024.

[21] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. Viton: An image-based virtual try-on network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7543–7552, 2018.

[22] Xintong Han, Xiaojun Hu, Weilin Huang, and Matthew R Scott. Clothflow: A flow-based model for clothed person generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10471–10480, 2019.

[23] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6546–6555, 2018.

[24] Sen He, Yi-Zhe Song, and Tao Xiang. Style-based global appearance flow for virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3470–3479, 2022.

[25] Zijian He, Peixin Chen, Guangrun Wang, Guanbin Li, Philip HS Torr, and Liang Lin. Wildvidfit: Video virtual try-on in the wild via image-based controlled diffusion models. In *European Conference on Computer Vision*, pages 123–139. Springer, 2024.

[26] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

[27] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[28] Jiahui Huang, Zan Gojcic, Matan Atzmon, Or Litany, Sanja Fidler, and Francis Williams. Neural kernel surface reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4369–4379, 2023.

[29] Zaiyu Huang, Hanhui Li, Zhenyu Xie, Michael Kampffmeyer, Xiaodan Liang, et al. Towards hard-pose virtual try-on via 3d-aware global correspondence learning. *Advances in Neural Information Processing Systems*, 35:32736–32748, 2022.

[30] Jianbin Jiang, Tan Wang, He Yan, and Junhui Liu. Clothformer: Taming video virtual try-on in all module. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10799–10808, 2022.

[31] Michael Kazhdan, Matthew Bolitho, and Hugues Hoppe. Poisson surface reconstruction. In *Proceedings of the fourth Eurographics symposium on Geometry processing*, 2006.

[32] Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for human vision models. In *European Conference on Computer Vision*, pages 206–228. Springer, 2025.

[33] Jeongho Kim, Guojung Gu, Minho Park, Sunghyun Park, and Jaegul Choo. Stableviton: Learning semantic correspondence with latent diffusion model for virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8176–8185, 2024.

[34] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[35] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.

[36] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3383–3393, 2021.

[37] Jiefeng Li, Siyuan Bian, Chao Xu, Zhicun Chen, Lixin Yang, and Cewu Lu. Hybrik-x: Hybrid analytical-neural inverse kinematics for whole-body mesh recovery. *arXiv preprint arXiv:2304.05690*, 2023.

[38] Peike Li, Yunqiu Xu, Yunchao Wei, and Yi Yang. Self-correction for human parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):3260–3271, 2020.

[39] Yifei Li, Hsiao-yu Chen, Egor Larionov, Nikolaos Sarafianos, Wojciech Matusik, and Tuur Stuyck. Diffavatar: Simulation-ready garment optimization with differentiable simulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4368–4378, 2024.

[40] Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. Common diffusion noise schedules and sample steps are flawed. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 5404–5411, 2024.

[41] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023.

[42] Heusel Martin, Ramsauer Hubert, Unterthiner Thomas, Nessler Bernhard, and Hochreiter Sepp. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30:6626–6637, 2017.

[43] Matiur Rahman Minar and Heejune Ahn. Cloth-vton: Clothing three-dimensional reconstruction for hybrid image-based virtual try-on. In *Proceedings of the Asian conference on computer vision*, 2020.

[44] Davide Morelli, Alberto Baldrati, Giuseppe Cartella, Marcella Cornia, Marco Bertini, and Rita Cucchiara. Ladi-vton: Latent diffusion textual-inversion enhanced virtual try-on. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 8580–8589, 2023.

[45] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[46] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023.

[47] Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024.

[48] Lingteng Qiu, Xiaodong Gu, Peihao Li, Qi Zuo, Weichao Shen, Junfei Zhang, Kejie Qiu, Weihao Yuan, Guanying Chen, Zilong Dong, et al. Lhm: Large animatable human reconstruction model from a single image in seconds. *arXiv preprint arXiv:2503.10625*, 2025.

[49] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[50] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[51] Zehong Shen, Huaijin Pi, Yan Xia, Zhi Cen, Sida Peng, Zechen Hu, Hujun Bao, Ruizhen Hu, and Xiaowei Zhou. World-grounded human motion recovery via gravity-view coordinates. In *SIGGRAPH Asia Conference Proceedings*, 2024.

[52] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J Black. Wham: Reconstructing world-grounded humans with accurate 3d motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2070–2080, 2024.

[53] Robert W Sumner, Johannes Schmid, and Mark Pauly. Embedded deformation for shape manipulation. In *ACM siggraph 2007 papers*, pages 80–es. 2007.

[54] Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, Liang Lin, and Meng Yang. Toward characteristic-preserving image-based virtual try-on network. In *Proceedings of the European conference on computer vision (ECCV)*, pages 589–604, 2018.

[55] Yuanbin Wang, Weilun Dai, Long Chan, Huanyu Zhou, Aixi Zhang, and Si Liu. Gpd-vvto: Preserving garment details

in video virtual try-on. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7133–7142, 2024.

[56] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[57] Zhenzhen Weng, Jingyuan Liu, Hao Tan, Zhan Xu, Yang Zhou, Serena Yeung-Levy, and Jimei Yang. Template-free single-view 3d human digitalization with diffusion-guided lrm. *arXiv preprint arXiv:2401.12175*, 2024.

[58] Zhenyu Xie, Zaiyu Huang, Xin Dong, Fuwei Zhao, Haoye Dong, Xijin Zhang, Feida Zhu, and Xiaodan Liang. Gp-vton: Towards general purpose virtual try-on via collaborative local-flow global-parsing learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23550–23559, 2023.

[59] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J. Black. ICON: Implicit Clothed humans Obtained from Normals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13296–13306, 2022.

[60] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J. Black. ECON: Explicit Clothed humans Optimized via Normal integration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[61] Yuhao Xu, Tao Gu, Weifeng Chen, and Chengcai Chen. Ootdiffusion: Outfitting fusion based latent diffusion for controllable virtual try-on. *arXiv preprint arXiv:2403.01779*, 2024.

[62] Zhengze Xu, Mengting Chen, Zhao Wang, Linyu Xing, Zhonghua Zhai, Nong Sang, Jinsong Lan, Shuai Xiao, and Changxin Gao. Tunnel try-on: Excavating spatial-temporal tunnels for high-quality virtual try-on in videos. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 3199–3208, 2024.

[63] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1481–1490, 2024.

[64] Hongwen Zhang, Yating Tian, Yuxiang Zhang, Mengcheng Li, Liang An, Zhenan Sun, and Yebin Liu. Pymaf-x: Towards well-aligned full-body model regression from monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12287–12303, 2023.

[65] Jinsong Zhang, Kun Li, Yu-Kun Lai, and Jingyu Yang. Pise: Person image synthesis and editing with decoupled gan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7982–7990, 2021.

[66] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[67] Zechuan Zhang, Zongxin Yang, and Yi Yang. Sifu: Side-view conditioned implicit function for real-world usable clothed human reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9936–9947, 2024.

[68] Fuwei Zhao, Zhenyu Xie, Michael Kampffmeyer, Haoye Dong, Songfang Han, Tianxiang Zheng, Tao Zhang, and Xiaodan Liang. M3d-vton: A monocular-to-3d virtual try-on network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13239–13249, 2021.

[69] Jun Zheng, Fuwei Zhao, Youjiang Xu, Xin Dong, and Xiaodan Liang. Viton-dit: Learning in-the-wild video try-on from human dance videos via diffusion transformers. *arXiv preprint arXiv:2405.18326*, 2024.

[70] Xiaojing Zhong, Zhonghua Wu, Taizhe Tan, Guosheng Lin, and Qingyao Wu. Mv-ton: Memory-based video virtual try-on network. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 908–916, 2021.

[71] Jingkai Zhou, Benzhi Wang, Weihua Chen, Jingqi Bai, Dongyang Li, Aixi Zhang, Hao Xu, Mingyang Yang, and Fan Wang. Realisdance: Equip controllable character animation with realistic hands. *arXiv preprint arXiv:2409.06202*, 2024.

[72] Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Zilong Dong, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. Champ: Controllable and consistent human image animation with 3d parametric guidance. In *European Conference on Computer Vision*, pages 145–162. Springer, 2024.

3DV-TON: Textured 3D-Guided Consistent Video Try-on via Diffusion Models

Supplementary Material

A. HR-VVT benchmark.

Owing to the limitations of the exist datasets, which exhibits relatively simple scenarios, it is challenging to accurately assess video try-on methods. Therefore, we have constructed a high-resolution (˜720p) video try-on benchmark called HR-VVT that includes 130 videos with 50 upper-body clothing, 40 lower-body clothing, and 40 dresses, with a variety of garments and motions in complex scenarios. Figure 14 show some examples in our benchmark.

Our HR-VVT benchmark was sourced from e-commerce platforms for research purposes and will remain strictly reserved for academic use. Our framework contains no personal identity information, with facial regions excluded from inpainting operations to ensure privacy preservation during the training process.

B. Disccusion

Limitation. Although we have significantly reduced the reconstruction time of clothed 3D humans by improving the optimization objectives of SMPL refinement and keeping it within an acceptable inference time, this is still insufficient in scenarios with higher speed requirements. Recently, works [48] on reconstructing animatable clothed 3D humans using a single feed-forward approach has greatly accelerated inference times and achieved remarkable improvements in visual quality. We believe that updating our 3D guidance pipeline to a single feed-forward paradigm can accelerate the reconstruction process, further advancing the application of textured 3D human guidance in more scenarios.

Potential societal impact. This paper delves into the realm of video try-on generation. Because of the powerful generative capacity, these models pose risks such as the potential for misinformation and the creation of fake videos. We sincerely remind users to pay attention to generated content. Besides, it is crucial to prioritize privacy and consent, as generative models frequently rely on vast datasets that may include sensitive information. Users must remain vigilant about these considerations to uphold ethical standards in their applications. Note that our method only focus on technical aspect. Both videos and model weights used in this paper will be open-released.

C. Animatable Textured 3D Guidance

Refine SMPL-X. Since our clothed human reconstruction method is based on the SMPL-X [41, 45] model, it is important to accurately align the estimated body and clothing silhouette. In practice, human pose and shape (HPS) regressors [36, 37, 51] can not give pixel-aligned SMPL-X fits. We refine the SMPL-X parameters by minimizing $\mathcal{L}_{\text{SMPL-X}}$ in Section 3.1 of our paper.

Unlike the optimization of shape β, pose θ, and translation t of SMPL-X in ICON [59], Since we primarily focus on the clothing area and do not have high accuracy requirements for the body pose details, we adjust the optimization target without optimizing the SMPL pose θ. This allows us to significantly reduce the number of optimization steps to reduce the reconstruction time (˜30s). And we optimize the shape β, camera scale s, and translation t to mitigate the anomalies in loss caused by incomplete human body parts and inaccurate camera estimation in real data, which may lead to errors in the refined SMPL-X. We additionally introduce a unidirectional regularization penalty to prevent the incorrect decrease in loss caused by abnormal reduction in camera scale caused by partial bodies present in the training data.

As shown in Figure 15, if we optimize the pose of SMPL-X in Panel (a), the pose refinement may be abnormal due to the incomplete human body parts, leading to reconstruction failure. Thanks to the powerful generative capabilities of the diffusion model [27, 49], which do not require high precision for the pose accuracy in 3D guidance, we choose to freeze the pose parameters θ, as this approach is sufficient to yield usable results for the robust 3D guidance reconstruction. In Panel (b) and (c), current HPS regressors often yield inaccurate camera scale estimations. To address this issue, we simultaneously optimize the camera scale s applied to SMPL-X. However, for incomplete human bodies, the camera scale s tends to be abnormally reduced. We use a unidirectional regularization penalty to constrain the optimization direction of the s. Figure 19 demonstrates that our 3D pipeline is applicable to most scenarios.

Animation. To ensure smooth animation of the 3D guidance, we employ a video-based SMPL estimation approach [51, 52]. However, during the reconstruction phase, our input is an image, and to achieve more accurate reconstruction, we employ an image-based SMPL-X estimation method. Since there are differences in shape and other parameters between the body estimations from video and image-based methods, directly using the body sequences estimated from video for animating may result in texture distortion and deformation, as shown in Figure 16 (left). To address this issue, we utilize the video-based estimated SMPL for rigging the reconstructed clothed human before animation. Figure 16 (right) demonstrates that our method effectively avoids texture distortion and deformation.

Figure 14. **Illustration of the HR-VVT benckmark.**

D. Network Architecture

Temporal attention. Since our textured 3D guidance provides sufficiently explicit frame-level references, we find that our 3DV-TON can maintain texture consistency even when temporal attention is freezed (initialized with AnimateDiff [20]), albeit with some minor jitter and mask artifacts. Texture errors occur only when the 3D guidance hard to provide texture references, as shown in the first two columns of Figure 17. This demonstrates that our 3DV-TON, using textured 3D guidance, is capable of generating consistent texture motion rather than overly focusing on smoothing inconsistent content between frames.

E. More Results

As shown in Figure 18, our method can handle various shape, materials, and complex textures of clothing, while generating consistent texture motions.

For more qualitative comparisons and video try-on results, please refer to the project page.

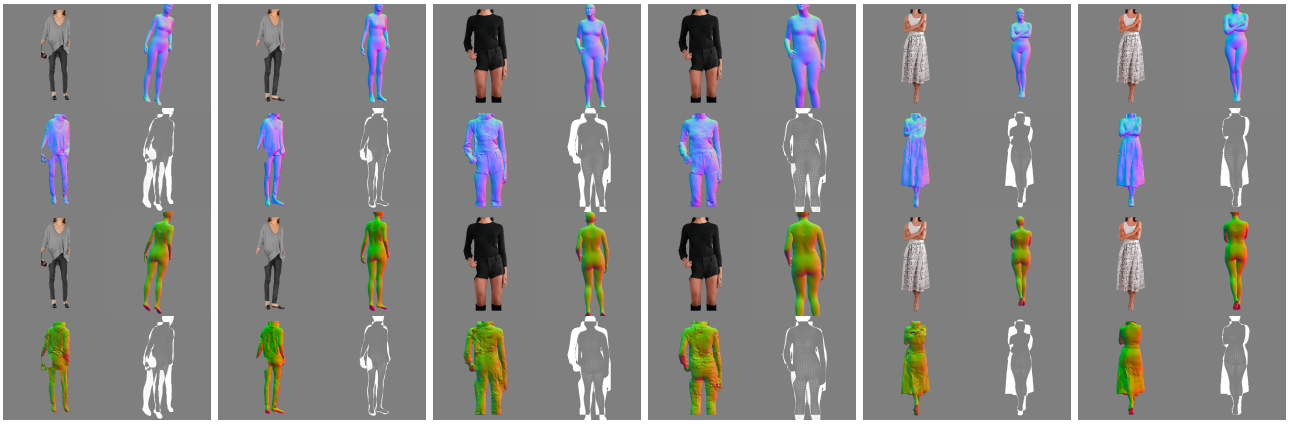

Figure 15. **Effectiveness of our SMPL-X Refinement.**

Figure 16. **Effectiveness of our textured 3D animation method.**

Figure 17. **Effectiveness of freezing temporal attention of our 3DV-TON.**

Figure 18. **More results generated by 3DV-TON.**

Figure 19. **Animated 3D guidance.**